Computational Linguistics and Chinese Language Processing
Vol. 17, No. 1, March 2012, pp. 43-64

The Polysemy Problem, an Important Issue in a Chinese to Taiwanese TTS System

Ming-Shing Yu* and Yih-Jeng Lin+

Abstract

This paper brings up an important issue, polysemy problems, in a Chinese to Taiwanese TTS (text-to-speech) system. Polysemy means there are words with more than one meaning or pronunciation, such as "我們" (we), "不" (no), "你" (you), "我" (I), and "要" (want). We first will show the importance of the polysemy problem in a Chinese to Taiwanese (C2T) TTS system. Then, we will propose some approaches to a difficult case of such problems by determining the pronunciation of "我們" (we) in a C2T TTS system. There are two pronunciations of the word "我們" (we) in Taiwanese, */ghun/* and */lan/*. The corresponding Chinese words are "阮" (we_1) and "咱" (we_2). We propose two approaches and a combination of the two to solve the problem. The results show that we have a 93.1% precision in finding the correct pronunciation of the word "我們" (we). Compared to the results of the layered approach, which has been shown to work well in solving other polysemy problems, the results of the combined approach are an improvement.

Keywords: Polysemy, Taiwanese, Chinese to Taiwanese TTS System, Layered Approach

1. Introduction

Besides Mandarin, Taiwanese is the most widely spoken dialect in Taiwan. According to Liang *et al.* (2004), about 75% of the population in Taiwan speaks Taiwanese. Currently, it is government policy to encourage people to learn one's mother tongue in schools because local languages are a part of local culture.

Researchers (Bao *et al.*, 2002; Chen *et al.*, 1996; Lin *et al.*, 1998; Lu, 2002; Shih *et al.*, 1996; Wu *et al.*, 2007; Yu *et al.*, 2005) have had outstanding results in developing Mandarin

* Department of Computer Science and Engineering, National Chung-Hsing University, Taichung 40227, Taiwan.
+ Department of Information Management, Chien-Kuo Technology University, Chang-hua 500, Taiwan.
E-mail: yclin@ctu.edu.tw

text-to-speech (TTS) systems over the past ten years. Other researchers (Ho, 2000; Huang, 2001; Hwang, 1996; Lin *et al.*, 1999; Pan & Yu, 2008; Pan, Yu, & Tsai, 2008; Yang, 1999; Zhong, 1999) have just begun to develop Taiwanese TTS systems. There are no formal characters for Taiwanese, so Chinese characters are officially used in Taiwan. Consequently, many researchers have focused on Chinese to Taiwanese (C2T) TTS systems. This means that the input of a so-called Taiwanese TTS system is Chinese text. Yang (1999) developed a method based on machine translation to help solve this problem. Since there are differences between Mandarin and Taiwanese, a C2T TTS system should have a text analysis module that can solve problems specific to Taiwanese. For instance, there is only one pronunciation for "我們" (we) in Chinese, but there are two pronunciations for "我們" (we) in Taiwanese.

Figure 1 shows a common structure of a C2T TTS system. In general, a C2T TTS system should contain four basic modules. They are (1) a text analysis module, (2) a tone sandhi module, (3) a prosody generation module, and (4) a speech synthesis module. A C2T TTS system also needs a text analysis module like that of a Mandarin TTS system. This module requires a well-defined bilingual lexicon. We also find that text analysis in a C2T TTS system should have functions not found in a Mandarin TTS system, such as phonetic transcription, digit sequence processing (Liang *et al.*, 2004), and a method for solving the polysemy problem. Solving the polysemy problem is the most complex and difficult of these. There has been little research on solving the polysemy problem. Polysemy means that a word has two or more meanings, which may lead to different pronunciations. For example, the word "他" (he) has two pronunciations in Taiwanese, */yi/* and */yin/*. The first pronunciation */yi/* of "他" (he) means "he," while the second pronunciation /yin/ of "他" (he) means "second-person possessive". The correct pronunciation of a word affects the comprehensibility and fluency of Taiwanese speech.

Many researchers have studied C2T TTS systems (Ho, 2000; Huang, 2001; Hwang, 1996; Lin *et al.*, 1999; Pan & Yu, 2008; Pan, Yu, & Tsai, 2008; Yang, 1999; Zhong, 1999). Nevertheless, none of the researchers considered the polysemy problem in a C2T TTS system. We think that solving the polysemy problem in a C2T TTS system is a fundamental task. The correct meaning of the synthesized words cannot be determined if this problem is not solved properly.

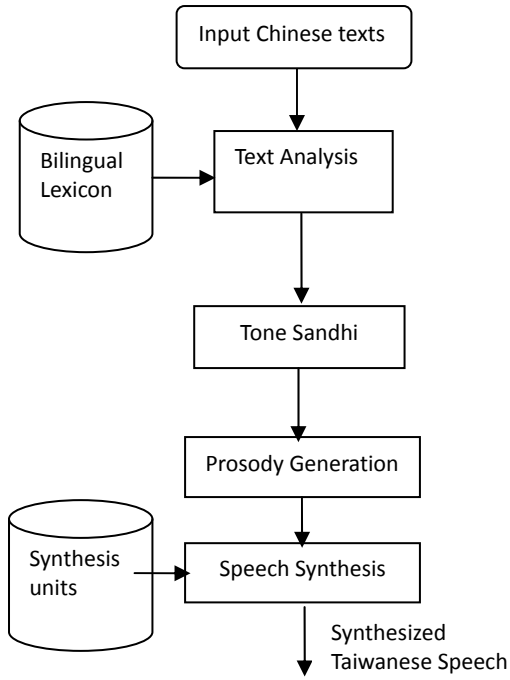

Figure 1. A Common module structure of a C2T TTS System.

The remainder of this paper is organized as follows. In Section 2, we will describe the polysemy problem in Taiwanese. We will give examples to show the importance of solving the polysemy problem in a C2T TTS system. Determining the correct pronunciation of the word "我們" (we) is the focus of the challenge in these cases. Section 3 is the description of the layered approach, which has been shown to work well in solving the polysemy problem (Lin *et al.*, 2008). Lin (2006) has also shown that the layered approach works very well in solving the polyphone problem in Chinese. We will apply the layered approach in determining the pronunciation of "我們" (we) in this section. In Section 4 and Section 5, we use two models to determine the pronunciation of the word "我們" (we) in sentences. The first approach in Section 4 is called the word-based unigram model (WU). The second approach, which will be applied in Section 5, is the word-based long-distance bigram model (WLDB). We also make some new inferences in these two sections. Section 6 shows a combination of the two models discussed in Section 4 and Second 5 for a third approach to solving the polysemy problem. Finally, in Section 7, we summarize our major findings and outline some future works.

2. Polysemy Problems in Taiwanese

Unlike in Chinese, the polysemy problem in Taiwanese appears frequently and is complex. We will give some examples to show the importance of solving the polysemy problem in a C2T TTS system.

The first examples feature the pronouns “你” (you), “我” (I), and “他” (he) in Taiwanese. These three pronouns have two pronunciations, each of which corresponds to a different meaning. Example 2.1 shows the pronunciations of the word “我” (I) and “你” (you) in Taiwanese. The two pronunciations of “我” (I) are */ghua/* with the meaning of “I” or “me” and */ghun/* with the meaning of “my”. The two pronunciations of “你” (you) are /*li*/ with the meaning of “you” and /*lin*/ with the meaning of “your”. If one chooses the wrong pronunciation, the utterance will carry the wrong meaning.

Example 2.1 我/ghua/過一會兒會拿幾本有關台語文化的書到你/*lin*/家給你/*li*/，你/*li*/可以不必到我/*ghun*/家來找我/*ghua*/拿。 (I will bring some books about Taiwanese culture to your house for you later; you need not come to my home to get them from me.)

Example 2.2 shows the two different pronunciations of “他” (he). They are /*yi*/, with the meaning of “he” or “him,” and /*yin*/, with the meaning of “his”.

Example 2.2 我看到他/yi/拿一盆蘭花回他/*yin*/家給他/*yin*/爸爸。 (I saw him bring an orchid back to his home for his father.)

The following examples focus on “不” (no), which has six different pronunciations. They are /*bho*/, /*m*/, /*bhei*/, /*bhuaih*/, /*mai*/, and /*but*/. Examples 2.3 through 2.6 show four of the six pronunciations.

Example 2.3 一般人並不/*bho*/容易看出它的重要性。 (It is not easy for a person to see its importance.)

Example 2.4 不/m/知浪費了多少國家資源。(We do not know how many national resources were wasted.)

Example 2.5 讓人聯想不/*bhei*/到他與機械的關係。 (One would not come to the proper conclusion regarding the relationship between that person and machines.)

Example 2.6 華航使用之航空站交通已不/*but*/如從前方便。 (The traffic at the airport is not as convenient as it was in the past for China Airlines.)

Examples 2.7 through 2.9 are examples of pronunciations of the word “上” (up). The word “上” (up) has three pronunciations. They are /*ding*/, /*siong*/, and /*jiunn*/. The meaning of the word “上” (up) in Example 2.7 has the sense of “previous”. Example 2.8 shows a case where “上” (up) means “on”. Example 2.9 is an example of the use of “上” (up) to mean, “get on”.

Example 2.7 我上/*ding*/個月花了好多錢去買有關台語的教科書。(Last month, I spent so much money on buying Taiwanese textbooks.)

Example 2.8 我是在這地圖上/*siong*/的哪裡？ (Where am I on this map?)

Example 2.9 我上/*jiunn*/了公車後才發現我搭錯車了。 (After I got on the bus, I realized that I boarded the wrong one.)

Another word we want to discuss is "下" (down). The word "下" (down) has four pronunciations. They are /*ha*/, /*ao*/, /*loh*/, and /*ei*/. Examples 2.10–2.13 are some examples of pronunciations of the word "下" (down). The meaning of "下" (down) in Example 2.10 is "close" or "end". Example 2.11 shows how the same word can mean "next". Example 2.12 illustrates the meaning "falling". Example 2.13 shows another example of it used to mean "next".

Example 2.10 我今天將在十點下/*ha*/課。 (I will finish my class at ten o'clock today.)

Example 2.11 台中下/*ao*/星期有甚麼音樂會？ (What concerts are scheduled for next week in Taichung?)

Example 2.12 彰化已經開始下/*loh*/大雨了。 (It has begun to rain heavily in Changhua.)

Example 2.13 請問下/*ei*/一列火車何時開出？ (Excuse me. Could you please tell me when the next train will depart?)

We have proposed a layered approach in predicting the pronunciations "上" (up), "下" (down), and "不" (no) (Lin *et al.*, 2008). The layered approach works very well in solving the polysemy problems in a C2T TTS system. A more difficult case of the polysemy problem will be encountered in this paper.

In addition to the above words, another difficult case is "我們" (we). Taiwanese speakers arrive at the correct pronunciation of the word "我們" (we) by deciding whether to include the listener in the pronoun.

Unlike Chinese, "我們" (we) has two pronunciations with different meanings when used in Taiwanese. This word can include (1) both the speaker and listener(s) or (2) just the speaker. These variations lead to two different pronunciations in Taiwanese, */lan/* and */ghun/*. The Chinese characters for */lan/* and */ghun/* are "咱" (we) and "阮" (we), respectively. The following example helps to illustrate the different meanings. More examples to illustrate these differences will be used later in this section.

Assume first that Jeffrey and his younger brother, Jimmy, ask their father to take them to see a movie then go shopping. Jeffrey can say the following to his father:

Example 2.14 爸爸你要記得帶我們一起去看電影，我們看完電影後，再一起去逛街。 (Daddy, remember to take us to see a movie and go shopping with us after we see the movie.)

The pronunciation of the first word “我們” (we) in Example 2.14 is */ghun/* in Taiwanese since the word “我們” (we) does not include the listener, Jeffrey’s father. The second instance of “我們” (we), however, is pronounced /*lan*/ since this instance includes both the speaker and the listener.

The pronunciation of “我們” (we) in Example 2.15 is */ghun/* in Taiwanese since the word “我們” (we) includes Jeffrey and Jimmy but does not include the listener, Jeffrey’s father.

Example 2.15 爸爸，我要和弟弟去看電影，我們看完電影後，會一起去逛街。 (Daddy, I will go to see a movie with my younger brother, and the two of us will go shopping after seeing the movie.)

If a C2T TTS system cannot identify the correct pronunciation of the word “我們” (we), we cannot understand what the synthesized Taiwanese speech means. In a C2T TTS system, it is necessary to decide the correct pronunciation of the Chinese word “我們” (we) in order to have a clear understanding of synthesized Taiwanese speech.

Distinguishing different kinds of meanings of “我們” (we) is a semantic problem. It is a difficult but important issue to be overcome in the text analysis module of a C2T TTS system. As there is only one pronunciation of “我們” (we) in Mandarin, a Mandarin TTS system does not need to identify the meaning of the word “我們” (we).

To compare this work with the research in Hwang *et al.* (2000) and Yu *et al.* (2003), determining the meaning of the word “我們” (we) may be more difficult than solving the non-text symbol problem. A person can determine the relationship between the listeners and the speaker then determine the meaning of the word “我們” (we). It is more difficult, however, for a computer to recognize the relationship between the listeners and speakers in a sentence.

Since determining whether listeners are included is a context-sensitive problem, we need to look at the surrounding words, sentences, or paragraphs to find the answer.

Let us examine the following Chinese sentence (Example 2.16) to help clarify the problem.

Example 2.16 我們必須加緊腳步改善台北市的交通狀況。 (We should press forward to improve the traffic of Taipei City.)

It is difficult to determine the Taiwanese pronunciation of the word “我們” (we) in Example 2.16 from the information in this sentence. To get the correct pronunciation of the word “我們” (we), we need to expand the sentence by adding words to the subject, *i.e.*, look forward, and predicate, *i.e.*, look backward. Assume that, when we add words to the subject and the predicate, we have a sentence that looks like Example 2.17:

Example 2.17 台北市長馬英九在接見美國記者時指出:「我們必須加緊腳步改善台北市的交通狀況。」 (Taipei city mayor Ma Ying-Jeou said that we should press

forward to improve the traffic of Taipei city when he received some reporters from the USA.)

As the reporters from the USA have no obligation to improve the traffic of Taipei, we can conclude that "我們" (we) does not include them. Therefore, it is safe to say that the correct pronunciation of the word "我們" (we) in Example 2.17 should be */ghun/*.

On the other hand, if the sentence reads as in Example 2.18 and context is included, the pronunciation of the word "我們" (we) should be */lan/*. We can find some important keywords such as "台北市長" (the Taipei city mayor) and "市府會議" (a meeting of the city government).

Example 2.18 台北市長馬英九在市府會議中指出:「我們必須加緊腳步改善台北市的交通狀況。」 (In a meeting of the city government, the Taipei city mayor, Ma Ying-Jeou, said that we should press forward to improve the traffic of Taipei City.)

When disambiguating the meaning of some non-text symbols, such as "/", ":", and "-" the keywords to decide the pronunciation of the special symbols may be within a fixed distance from the given symbol. Nevertheless, the keywords can be at any distance from the word "我們" (we), as per Example 2.19. Some words that could be used to determine the pronunciation of "我們" (we), such as "市府會議" (a meeting of the city government), "台北市長" (the Taipei city mayor), and "馬英九" (Ma Ying-Jeou), are at various distances from "我們" (we).

Example 2.19 在今天的市府會議中，台北市長馬英九提到關於台北市的交通問題時，馬市長說:「我們必須加緊腳步改善台北市的交通狀況。」 (In a meeting of the city government, the Taipei city mayor, Ma Ying-Jeou, talked about the problem of the traffic in Taipei city. Mayor Ma said that we should press forward to improve the traffic of Taipei city.)

These examples illustrate the importance of determining the proper pronunciation for each word in a C2T TTS system. Compared to other cases of polysemy, determining the proper pronunciation of the word "我們" (we) in Taiwanese is a difficult task. We will focus on solving the polysemy problem of the word "我們" (we) in this paper.

3. Using the Layered Approach to Determine the Pronunciation of "我們" (we)

Lin (2006) showed that the layered approach worked very well in solving the polyphone problem in Chinese. Lin (2006) also showed that using the layered approach to solve the polyphone problem is more accurate than using the CART decision tree. We also show that using the layered approach in solving the polysemy problems of other words has worked well

in our research (Lin *et al*., 2008). We will apply the layered approach in solving the polysemy problem of “我們” (we) in Taiwanese.

3.1 Description of Experimental Data

First, we will describe the experimental data used in this paper. The experimental data is comprised of over forty thousand news items from eight news categories, in which 1,546 articles contain the word “我們” (we). The data was downloaded from the Internet from August 23, 2003 to October 21, 2004. The distribution of these articles is shown in Table 1. We determined the pronunciation of each “我們” (we) manually.

Table 1. Distribution of experimental data

News Category	Number of News Items	Number of News Items Containing the word "我們"	Percentage
International News	2242	326	14.5%
Travel News	9273	181	1.9%
Local News	6066	95	1.5%
Entertainment News	3231	408	12.6%
Scientific News	3520	100	2.8%
Social News	4936	160	3.2%
Sports News	2811	193	6.9%
Stock News	8066	83	1.0%
Total Number of News Items	40145	1546	3.9%

As shown in Table 2, in the 1,546 news articles, “我們” occurred 3,195 times. In our experiment, 2,556 samples were randomly chosen for the training data while the other 639 samples were added to the test data. In the training data, there were 1,916 instances with the pronunciation of */ghun/* for the Chinese character “阮” and 640 instances with the pronunciation of */lan/* for the Chinese character “咱”.

Table 2. Distribution of training and testing data.

Frequency of “我們”	Pronunciation */lan/*	Pronunciation */ghun/*	Total Frequency
Training data	640	1,916	2,556
Test data	160	479	639
Token frequency of “我們”	800	2,395	3,195

3.2 Description of Layered Approach

Figure 2 shows the layered approach to the polysemy problem with an input test sentence. We use Example 3.1 to illustrate how the layered approach works.

Example 3.1 爸爸 告訴 我們 過 馬路 要 小心。 (Dad told us to be careful when crossing the street.)

Example 3.1 is an utterance in Chinese with segmentation information. Spaces were used to separate the words in Example 3.1. We want to predict the correct pronunciation for the word "我們" (we) in Example 3.1.

As depicted in Figure 2, there are four layers in our approach. We set ($w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2}$) as (爸爸,告訴,我們,過,馬路). This pattern (爸爸,告訴,我們,過,馬路) will be the input for Layer 4. Nevertheless, as this pattern is not found in the training data, we cannot decide the pronunciation of "我們" (we) with this pattern. We then use two patterns ($w_{-2}, w_{-1}, w_0, w_{+1}$) and ($w_{-1}, w_0, w_{+1}, w_{+2}$) to derive (爸爸,告訴,我們,過) and (告訴,我們,過,馬路), respectively, as the inputs for Layer 3. Since we cannot find any patterns in the training data that match either of these patterns, the pronunciation cannot be decided in this layer.

Three patterns are used in Layer 2. They are (爸爸,告訴,我們), (告訴,我們,過), and (我們,過,馬路). We find that the pattern (爸爸,告訴,我們) has appeared in training data. The frequencies are 2 for pronunciation */ghun/* and 1 for */lan/*. Thus, the probabilities for the possible pronunciations of "我們" (we) in Example 3.1 are 2/3 for */ghun/* and 1/3 for */lan/*. We can conclude that the predicted pronunciation is */ghun/*. The layered approach terminates in Layer 2 in this example. If the process did not terminate prematurely, as in this example, it would have terminated in Layer 1, as shown by the dashed lines in Figure 2.

3.3 Results of Using the Layered Approach

We used the experimental data mentioned in 3.1. There are 3,159 samples in the corpus. We used 2,556 samples to train the four layers. The other 639 samples form the test data. Table 3 shows the accuracy of using the layered approach based on word patterns. Thus, the features in the layered approach are words. The results show that the layered approach does not work well. The overall accuracy is 77.00%.

Table 3. Results of using the layered approach with word pattern.

	Number of test samples	Number of correct samples	Accuracy rate
/ghun/	479	445	92.90%
/lan/	160	47	29.38%
Total	639	492	77.00%

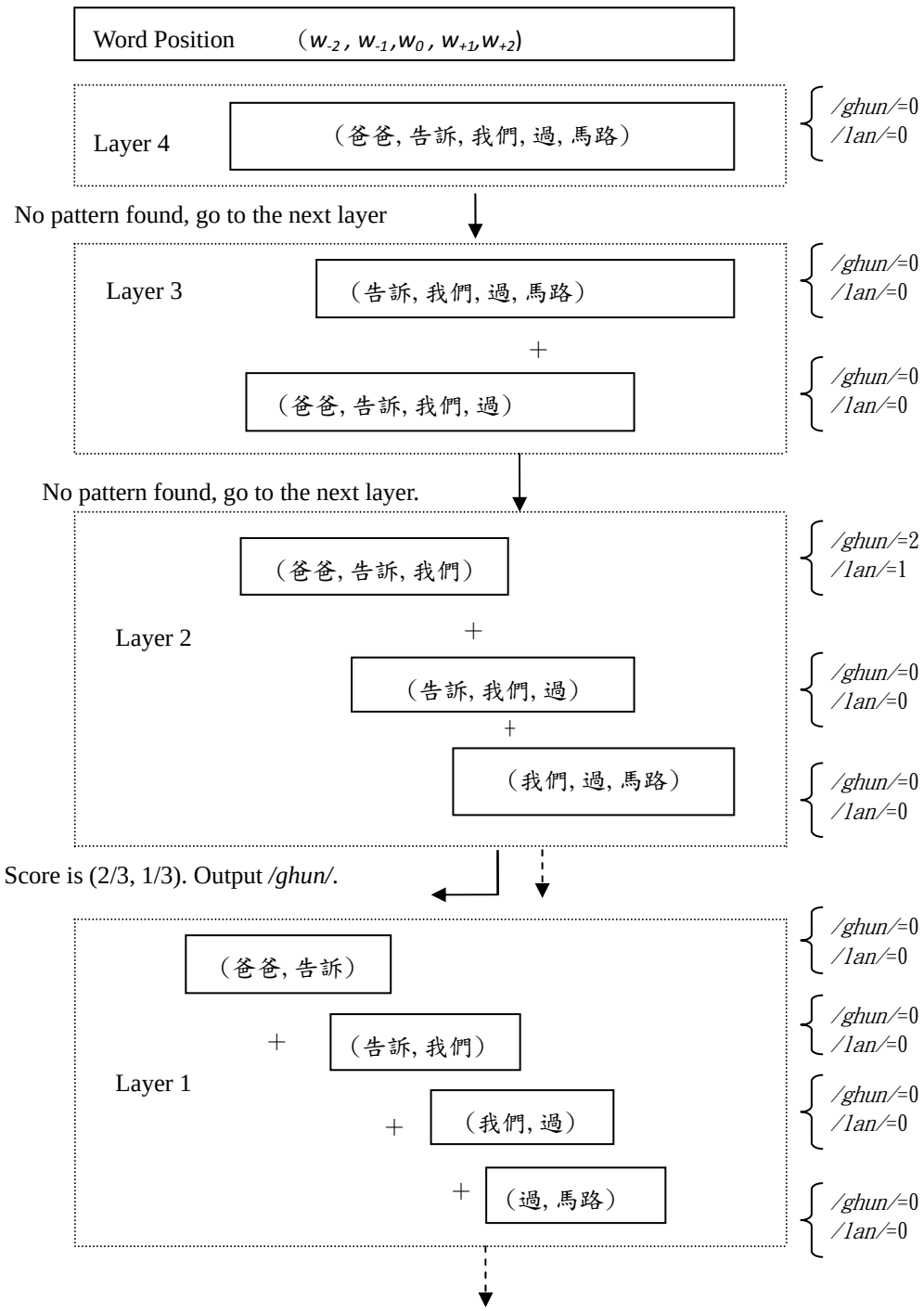

Figure 2. An example applying the layered approach.

4. Word-based Unigram Language Model

In this section, we propose a word-based unigram language model (WU). Two statistical results are needed in this model. Statistical results were compiled for (1) the frequency of appearance for words that appear to the left of "我們" (we) in the training data and (2) the frequencies for words that appear to the right. Each punctuation mark was treated as a word. Each testing sample looks like the following:

$$w_{-M}\, w_{-(M-1)} \ldots w_{-2}\, w_{-1}\ \boxed{\text{我們}}\ w_{+1}\, w_{+2} \ldots w_{+(N-1)}\, w_{+N}$$

where w_{-i} is the i^{th} word to the left of "我們" (we) and w_i is the i^{th} word to the right. The following formulae were used to find four different scores for each testing sample: $S_{uL}(/lan/)$, $S_{uR}(/lan/)$, $S_{uL}(/ghun/)$, and $S_{uR}(/ghun/)$.

$$S_{uL}(/lan/) = \sum_{j=1}^{M} \frac{\dfrac{C(/lan/\,\&\,w_{-j})}{T_{uL}(/lan/)}}{\dfrac{C(/lan/\,\&\,w_{-j})}{T_{uL}(/lan/)} + \dfrac{C(/ghun/\,\&\,w_{-j})}{T_{uL}(/ghun/)}} \tag{1}$$

$$S_{uR}(/lan/) = \sum_{j=1}^{N} \frac{\dfrac{C(/lan/\,\&\,w_{+j})}{T_{uR}(/lan/)}}{\dfrac{C(/lan/\,\&\,w_{+j})}{T_{uR}(/lan/)} + \dfrac{C(/ghun/\,\&\,w_{+j})}{T_{uR}(/ghun/)}} \tag{2}$$

$$S_{uL}(/ghun/) = \sum_{j=1}^{M} \frac{\dfrac{C(/ghun/\,\&\,w_{-j})}{T_{uL}(/ghun/)}}{\dfrac{C(/lan/\,\&\,w_{-j})}{T_{uL}(/lan/)} + \dfrac{C(/ghun/\,\&\,w_{-j})}{T_{uL}(/ghun/)}} \tag{3}$$

$$S_{uR}(/ghun/) = \sum_{j=1}^{N} \frac{\dfrac{C(/ghun/\,\&\,w_{+j})}{T_{uR}(/ghun/)}}{\dfrac{C(/lan/\,\&\,w_{+j})}{T_{uR}(/lan/)} + \dfrac{C(/ghun/\,\&\,w_{+j})}{T_{uR}(/ghun/)}} \tag{4}$$

where

$$T_{uL}(/lan/) = \sum_{l=1}^{uL} C(/lan/\,\&w_{-l}) \tag{5}$$

$$T_{uL}(/ghun/) = \sum_{p=1}^{uL} C(/ghun/\,\&w_{-p}) \tag{6}$$

$$T_{uR}(/lan/) = \sum_{l=1}^{uR} C(/lan/\,\&w_{+l}) \tag{7}$$

$$T_{uR}(/ghun/) = \sum_{p=1}^{uR} C(/ghun/\,\&w_{+p}) \tag{8}$$

uL different kinds of words appear on the left side of "我們" (we) in the training corpus. $T_{uL}(/lan/)$ is the total frequency of these *uL* words in the training data where the pronunciation of "我們" (we) is */lan/*. Similarly, $T_{uL}(/ghun/)$ represents the total frequency of *uL* words where "我們" (we) is pronounced */ghun/*. *uR* is the number of different words that appear to the right side of "我們" (we) in the training corpus. $T_{uR}(/lan/)$ and $T_{uR}(/ghun/)$ are the total frequencies of these *uR* words in the training data where pronunciation of "我們" (we) is */lan/ and /ghun/*, respectively. $C(/ghun/\&w_p)$ is the frequency that the word w_p appears in the training corpus where the pronunciation of "我們" (we) is */ghun /*. $\frac{C(/lan/\&w_{-j})}{T_{uL}(/lan/)}$ in (1) means the significance of pronunciation */lan/* of word w_{-j} in training data.

Formulae (1) through (4) were applied to each test sample to produce four scores. The scores were $S_{uL}(/lan/)$ for the words to the left of "我們" (we) when the pronunciation was */lan/*, $S_{uR}(/lan/)$ for the words to the right when the pronunciation was */lan/*, $S_{uL}(/ghun/)$ for the words to the left of "我們" (we) when the pronunciation was */ghun/*, and $S_{uR}(/ghun/)$ for the words to the right when the pronunciation was */ghun/*. The pronunciation of "我們" (we) is */lan/* if $S_{uL}(/lan/) + S_{uR}(/lan/) > S_{uL}(/ghun/) + S_{uR}(/ghun/)$. The result is */ghun/* otherwise.

The experiments were inside and outside tests. First, we applied WU with the training data mentioned in Section 3.1 to find the best ranges in determining the pronunciation of "我們" (we). We defined a window as (*M*, *N*), where *M* was number of words to the left of "我們" (we) and *N* was the number of words to the right. Three hundred and ninety nine (20*20-1=399) different windows were applied when using the WU model. As shown in Table 4, the best result from an inside test was 87.00%, with a window of (17, 10).

The best result when the correct pronunciation of "我們" (we) was */ghun/* was 94.01%, achieved when the window was (12, 6). Nevertheless, the results when the pronunciation was */lan/* and the window was the same were not good. The highest accuracy achieved was 45.48%. Also, as shown in 4th row of Table 4, the best result when applying WU when the pronunciation was */lan/* was just 77.88%, when the window was (19, 14). This shows that WU did not work well when the pronunciation of "我們" (we) was */lan/*.

Table 4. The results of the inside test of applying WU.

Window Size (*M*, *N*)	Accuracy when the pronunciation is */ghun/*	Accuracy when the pronunciation is */lan/*	Overall accuracy
(17, 10)	91.04%	74.92%	87.00%
(12,6)	94.01%	45.48%	81.85%
(19, 14)	88.75%	77.88%	86.03%

We applied WU with a window of (17, 10) for testing data. The overall accuracy of the outside tests was 75.59%. The accuracies were 90.40% and 31.25% when the pronunciations were */ghun/* and */lan/*, respectively.

5. Word-based Long Distance Bigram Language Model

We will bring up the word-based long-distance bigram language model (WLDB) in this section. According to Section 2 of this paper, there are two different meanings for "我們" (we). The two meanings are different in that one includes the listener(s) and the other does not. We propose a modification of the WU model by having two words appear together in the text to clarify the relationship between the speaker and listener(s). Examples of this modification are "台北市長" (the Taipei city mayor) and "美國記者" (the reporter(s) from the USA) in Example 2.17 and "台北市長" and "市府會議" (a city government meeting) in Examples 2.18 and 2.19.

For each testing sample,

$$w_{-M}\, w_{-(M-1)} \ldots w_{-2}\, w_{-1}\ \boxed{\text{我們}}\ w_{+1}\, w_{+2} \ldots w_{+(N-1)}\, w_{+N}\ .$$

The following formulae were used to find four scores for each testing sample, $S_{bL}(/lan/)$, $S_{bR}(/lan/)$, $S_{bL}(/ghun/)$, and $S_{bR}(/ghun/)$.

$$S_{bL}(/lan/) = \sum_{i=1}^{M}\sum_{j=i}^{M} \frac{\dfrac{C(/lan/ \,\&\, w_{-i} \,\&\, w_{-j})}{T_{bL}(/lan/)}}{\dfrac{C(/lan/ \,\&\, w_{-i} \,\&\, w_{-j})}{T_{bL}C(/lan/)} + \dfrac{C(/ghun/ \,\&\, w_{-i} \,\&\, w_{-j})}{T_{bL}C(/ghun/)}} \tag{9}$$

$$S_{bR}(/lan/) = \sum_{i=1}^{N}\sum_{j=i}^{N} \frac{\dfrac{C(/lan/ \,\&\, w_{i} \,\&\, w_{j})}{T_{bR}(/lan/)}}{\dfrac{C(/lan/ \,\&\, w_{+i} \,\&\, w_{+j})}{T_{bR}(/lan/)} + \dfrac{C(/ghun/ \,\&\, w_{+i} \,\&\, w_{+j})}{T_{bR}(/ghun/)}} \tag{10}$$

$$S_{bL}(/ghun/) = \sum_{i=1}^{M}\sum_{j=i}^{M} \frac{\dfrac{C(/ghun/ \,\&\, w_{-i} \,\&\, w_{-j})}{T_{bL}(/ghun/)}}{\dfrac{C(/ghun/ \,\&\, w_{-i} \,\&\, w_{-j})}{T_{bL}(/ghun/)} + \dfrac{C(/lan/ \,\&\, w_{-i} \,\&\, w_{-j})}{T_{bL}(/lan/)}} \tag{11}$$

$$S_{bR}(/ghun/) = \sum_{i=1}^{N}\sum_{j=i}^{N} \frac{\dfrac{C(/ghun/ \,\&\, w_{i} \,\&\, w_{j})}{T_{bR}(/ghun/)}}{\dfrac{C(/ghun/ \,\&\, w_{+i} \,\&\, w_{+j})}{T_{bR}(/ghun/)} + \dfrac{C(/lan/ \,\&\, w_{+i} \,\&\, w_{+j})}{T_{bR}(/lan/)}} \tag{12}$$

where

$$T_{bL}(/lan/) = \sum_{l=1}^{bL}\sum_{k=l}^{bL} C(/lan/ \,\&\, w_{-l} \,\&\, w_{-k}) \tag{13}$$

$$T_{bR}(/lan/) = \sum_{l=1}^{bR}\sum_{k=l}^{bR} C(/lan/ \,\&\, w_{+l} \,\&\, w_{+k}) \tag{14}$$

$$T_{bL}(/ghun/) = \sum_{l=1}^{bL}\sum_{k=l}^{bL} C(/ghun/ \,\&\, w_{-l} \,\&\, w_{-k}) \tag{15}$$

$$T_{bR}(/ghun/) = \sum_{l=1}^{bR} \sum_{k=l}^{bR} C(/ghun/ \& w_l \& w_k) \quad (16)$$

We assume that *bL* different words appear to the left of "我們" (we) in the training corpus and *bR* different words appear to the right. Formulae 9, 10, 11, and 12 were applied to each test sample, and they produced four scores. $C(/lan/\&w_i\&w_j)$ in (9) is the frequency at which words w_i and w_j appear in the training corpus when the pronunciation of "我們" (we) is */lan/*. $S_{bL}(/lan/)$ is the score for the words to the left of "我們" (we) when the pronunciation is */lan/*, and $S_{bR}(/lan/)$ is the score for the words to the right. Similarly, $S_{bL}(/ghun/)$ and $S_{bR}(/ghun/)$ represent the scores for the words to the left and right, respectively, when "我們" (we) is pronounced */ghun/*. In summary, the pronunciation of the word "我們" (we) is */lan/* if $S_{bL}(/lan/) + S_{bR}(/lan/) > S_{bL}(/ghun/) + S_{bR}(/ghun/)$. The pronunciation is */ghun/* otherwise.

We applied WLDB with the training data mentioned in Section 3.1 to find the best ranges in determining the pronunciation of "我們" (we). We defined a window of (*M*, *N*), where *M* was the number of words to the left and *N* was number of words to the right. Three hundred and sixty (19*19-1=360) different windows were applied in the analysis of using the WLDB model. As shown in the 2nd row of Table 5, the best result of the inside test was 94.25% with the best range being 11 words to the left of "我們" (we) and 7 words to the right.

The best result when the correct pronunciation of "我們" (we) was */lan/* was 99.87%, when the window was (11, 5). Nevertheless, the result for */ghun/* with the same window was not good. The highest accuracy achieved was 89.69%. As shown in the 3rd row of Table 5, the best result when applying WLDB when the pronunciation was */ghun/* was 93.48%, when the window was (4, 13). This shows that WLDB does not work well when the pronunciation of "我們" (we) is */ghun/*.

Table 5. The results of the inside test of applying WLDB.

Window Size (k_L, k_R)	Accuracy when the pronunciation is */ghun/*	Accuracy when the pronunciation is */lan/*	Overall accuracy
(11,7)	93.33%	97.04%	94.25%
(4, 13)	93.48%	93.61%	93.52%
(11,5)	89.69%	99.87%	92.15%

We applied the WLDB model to the test data using a window of (11, 7). The overall accuracy of outside tests was 85.72%. The accuracies were 83.26% and 93.10% when the pronunciations were */ghun/* and */lan/*, respectively.

6. The combined Approach

Based on the results from the two models, WU and WLDB, we can draw the following

conclusions: the word-based long distance bigram language model is good when the pronunciation is */lan/*, while the word-based unigram language model works well when the pronunciation is */ghun/*. In this section, we propose combining the models to achieve better results.

According to the inside experimental results shown in Table 4 and Table 5, we will combine the WU model with a window of (12, 6) and the WLDB model with a window of (11, 5) as our combined approach. This combination of WU and WLDB is similar to the approach used by Yu and Huang. We will try to find the possibility of making a correct choice when using WU or WLDB, which will be termed "confidence". We will adopt the output of the method with higher confidence.

6.1 Confidence Measure

The first step in this process is to find a confidence curve for each model. The goal is to estimate the confidence for each approach and assess the difference. The higher score is more likely to be the correct answer. To do so, we measure the accuracy of each division and use a regression to estimate the confidence measure.

Algorithm 1, below, will be used to find the confidence curve for the word-based unigram language model. As the total number of words in each input sample is not constant, we must first normalize the scores $Su_i(/lan/)$ and $Su_i(/ghun/)$. We will find the precision rates (PR_k) in the interval [0, 1] for $|NSu_i(/ghun/)- NSu_i(/lan/)|$ in Step 2 of Algorithm 1 for each i. We then find a regression curve for the PR_k. The regression curve is used to estimate the probability of making a correct decision when using WU. Therefore, it follows that, the higher the probability is, the greater the confidence we can have in the results from WU.

Algorithm 1: Finding the confidence curve of WU.

Input: The score for each training sample, $Su_i(/lan/)$ and $Su_i(/ghun/)$, where i=1,2,3, …, n and n is the number of training samples.

Output: A function for the confidence curve for the given $Su_i(/lan/)$ and $Su_i(/ghun/)$, i=1,2,3, …, n.

Algorithm:

Step 1: Normalize $Su_i(/lan/)$ and $Su_i(/ghun/)$ for each training sample i using the following formula:

$$NSu_i(/lan/)=Su_i(/lan/)/(\text{Total number of words in training sample } i)$$
$$NSu_i(/ghun/)=Su_i(/ghun/)/(\text{Total number of words in training sample } i)$$

Step 2: Let $d_i=| NSu_i(/ghun/)- NSu_i(/lan/)|$ and let $D=\{d_1, d_2,\ldots,d_n\}$. Find the accuracy rate for each interval using the following formula:

$$PR_k= C_k/N_k,\ k=1, 2, \ldots, 18$$

Here, C_k is the number of correct conjectures of training sample i with $(k\text{-}1)/18 \leqq d_i < (k+1)/18$, and N_k is the number of training sample i with $(k\text{-}1)/18 \leqq d_i < (k+1)/18$.

Step 3: Find a regression curve for PR_1, PR_2, …, PR_{18}. Output the function of the regression curve.

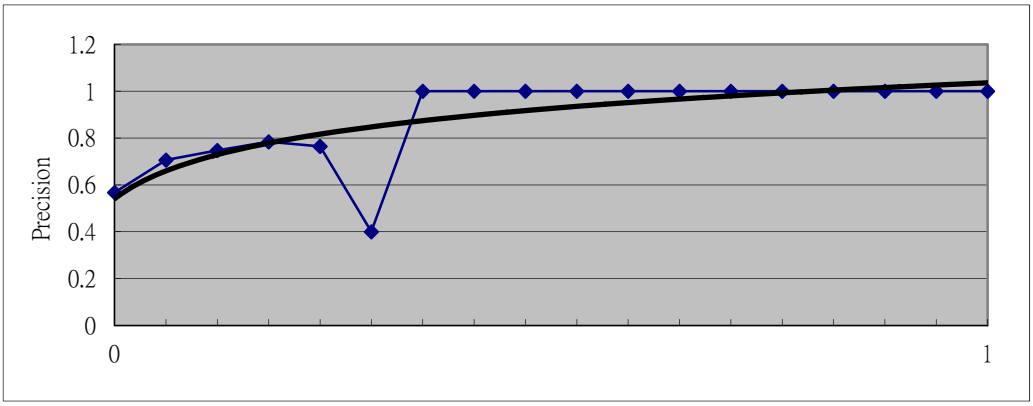

Figure 3. Estimate the confidence curve using WU. The function we attained is f(x)=0.1711*ln(x)+1.0357.

The confidence curve for WU is the black line in Figure 3. The function derived was $f(x)=0.1711*\ln(x)+1.0357$, where x is the absolute value of the difference between the normalized $Su_i(/lan/)$ and $Su_i(/ghun/)$.

Algorithm 2 is used to find the confidence curve for the word-based long-distance bigram language model (WLDB). We began by normalizing the scores of pronunciation $Sb_i(/lan/)$ and $Sb_i(/ghun/)$. In Step 2, we find the precision rates (PR_k) in the interval [0, 1] then calculate a regression curve for the PR_k. The regression curve will be used to estimate the probability of making a correct decision. Again, it follows that, the higher the probability, the more confidence in the results from using WLDB.

The confidence curve of WLDB is the black line in Figure 4, in which the function is $f(x) = 0.2346*\ln(x) + 1.0523$, where x is the difference between the normalized $Sp_i(/lan/)$ and $Sp_i(/ghun/)$.

Algorithm 2: Find the confidence curve of WLDB

Input: The score of each training sample, named $Sb_i(/lan/)$ and $Sb_i(/ghun/)$, where i=1, 2, 3, …, n, and n is the number of training samples.

Output: A function for the confidence curve for the given $Sb_i(/lan/)$ and $Sb_i(/ghun/)$, i=1, 2, 3, …, n.

Algorithm:

Step 1: Normalize $Sb_i(/lan/)$ and $Sb_i(/ghun/)$ for each training sample i using the following formula:

$$NSb_i(/lan/)=Sb_i(/lan/)/(\text{Total number of words in training sample } i)^2$$
$$NSb_i(/ghun/)=Sb_i(/ghun/)/(\text{Total number of words in training sample } i)^2$$

Step 2: Let $d_i=|\ NSb_i(/ghun/)- NSb_i(/lan/)|$ and let $D=\{d_1, d_2,\ldots,d_n\}$. Find the accuracy rate for each interval using the following formula:

$$PR_k= C_k/N_k,\ k=1, 2, \ldots, 13$$

where C_k is the number of correct conjectures of training samples i with $(k\text{-}1)/13 \leqq d_i < (k+1)/13$ and N_k is the number of training samples i with $(k\text{-}1)/13 \leqq d_i < (k+1)/13$.

Step 3: Find a regression curve for PR_1, PR_2, …, PR_{13}. Output the function of the regression curve.

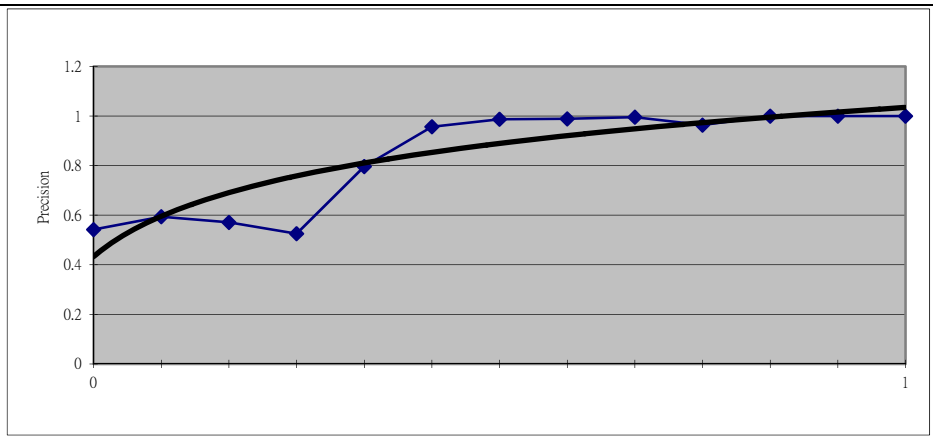

Figure 4. Estimate the confidence curve of WLDB. The function we attained is f(x)=0.2346*ln(x)+1.0523.

6.2 Determining the Pronunciation for "我們" (we)

After the functions for the confidence curves for the two models have been derived, the combined approach can be applied. The two models are used to determine the pronunciation of "我們" (we) for a given input text. The two functions for the confidence curves, derived in Section 6.1, are applied to evaluate the degree of confidence in the two models. Let the confidence curves of the two models be C_{WU} for WU and C_{WLDB} for WLDB. We will use the results obtained using WU under the condition $C_{WU} > C_{WLDB}$. Otherwise, we will use the results obtained from using the WLDB model.

Consider Figure 4, which is derived from the training data. The *x*-axis is the normalized difference between the two scores. The *y*-axis is the percentage of correct decisions. Take the example sentence "如果花旗希望繼續做我們的大股東，我們還是很歡迎". We want to predict the pronunciation of the first "我們" (we) in the above sentence. Its confidences were 0.875 for the WU model (choosing */ghun/*) and 0.761 for the WLDB model (choosing */lan/*). Since the confidence of the WU model was higher than that of the WLDB model, we adopted */ghun/* as the pronunciation.

6.3 Experimental Results Using Combined Models

We used the 639 testing samples described in Section 3.1. Among the 639 testing samples, there were 479 samples with the pronunciation */ghun/* and 160 samples with the pronunciation */lan/*.

We used the test data mentioned in 3.1 as the experimental data. The overall accuracy rate from applying the combined approach was 93.6%. The accuracy rate was 95.00% when the answer was */lan/*, and the accuracy rate was 93.1% when the answer was */ghun/*. Based on these results, it can be concluded that the combination of the two models works very well in determining the pronunciation of the word "我們" (we) for a given Chinese text.

The three approaches, WU, WLDB, and combined, are compared in Table 6. As shown in Table 6, the word-based long-distance bigram language model (WLDB) worked well in the case of */lan/* and achieved an accuracy rate of 93.10%. The word-based unigram language (WU) model worked well in the case of */ghun/* and achieved an accuracy rate of 90.40%. The combined approach, however, achieved higher accuracy rates in both cases, achieving accuracy as high as 93.6%.

Table 6. Comparison - WU, WLDB, and Combined approach.

	Accuracy using WU	Accuracy using WLDB	Accuracy combing the two models
/ghun/	90.40%	83.26%	93.10%
/lan/	31.25%	93.10%	95.00%
Total	75.59%	85.72%	93.60%

There is an important issue in the combined approach. When we use a language model like WLDB, we may encounter the problem of data scarcity. If data is scarce, the combined approach will use the result of the word-based unigram language model.

6.4 Discussion

Table 7 compares the accuracy of the approaches used in this paper. The findings show that the combined approach (CP) performed the best. We can conclude that layered approach does not work well in determining the pronunciation of "我們" (we) in Taiwanese. It also shows that the polysemy problem caused by "我們" (we) is more difficult and quite different from that caused by the words "上" (up), "下" (down), and "不" (no). This also shows that the viewpoints we gave in Section 2 are reasonable.

Table 7. A comparison of the proposed methods and the layered approach. CP refers to the combined approach, while LP refers to the layered approach. The combined approach achieved the highest accuracy.

	WU	WLDB	LP	CP
/ghun/	90.40%	83.26%	92.90%	93.10%
/lan/	31.25%	93.10%	29.38%	95.00%
Total	75.59%	85.72%	77.00%	93.60%

For our approaches, we might encounter the problem of data sparseness, especially with WLDB. It seems that this cannot be avoided in processing languages like Taiwanese, for which corpora are rare. We have tried to use part-of-speech information as the features in our approaches. The experimental results are not good. We also find that most cases can be solved by using WU or WLDB, and only about 5% are solved by using default values. This shows that our approach is suitable for the current data size. We have shown that our combined approach is promising.

7. Conclusion and Future Works

This paper proposes an elegant approach to determine the pronunciation of "我們" (we) in a C2T TTS system. Our methods work very well in determining the pronunciations of the Chinese word "我們" (we) in a C2T TTS system. Experimental results also show that the model used is better than the layered approach, the WU model, and the WLDB model. Polysemy problems in translating C2T are very common and it is imperative that they are solved in a C2T TTS system. We will continue to focus on other important polysemy problems in a C2T TTS system in the future.

The polysemy problem of "我們" (we) is more difficult than that of other words in Taiwanese. We have proposed a combined approach for this problem. If more training data can be prepared, the proposed approach can be expected to achieve better results. Nevertheless, as the training data needs to be processed manually, we will attempt to propose unsupervised approaches in the future.

To build a quality C2T TTS system is a long-term project because of the many issues in the text analysis phase. In contrast to a Mandarin TTS system, a C2T TTS system needs more textual analysis functions. In addition, two imperative tasks are the development of solutions for the polysemy problem and the tone sandhi problem.

Reference

Bao, H., Wang, A., & Lu, S. (2002). A Study of Evaluation Method for Synthetic Mandarin Speech, in *Proceedings of ISCSLP 2002, The Third International Symposium on Chinese Spoken Language Processing*, 383-386.

Chen, S. H., Hwang, S. H., & Wang, Y. R. (1996). A Mandarin Text-to-Speech System, *Computational Linguistics and Chinese Language Processing*, 1(1), 87-100.

Ho, C. C. (2000). *A Hybrid Statistical/RNN Approach to Prosody Synthesis for Taiwanese TTS*, Master thesis, Department of Communication Engineering, National Chiao Tung University.

Hunag, J. Y. (2001). *Implementation of Tone Sandhi Rules and Tagger for Taiwanese TTS*, Master thesis, Department of Communication Engineering, National Chiao Tung University.

Hwang, C. H. (1996). *Text to Pronunciation Conversion in Taiwanese*, Master thesis, Institute of Statistics, National Tsing Hua University.

Hwang, F. L., Yu, M. S., & Wu, M. J. (2000). The Improving Techniques for Disambiguating Non-Alphabet Sense Categories, in P*roceedings of ROCLING XIII*, 67-86.

Liang, M. S., Yang, R. C., Chiang, Y. C., Lyu, D. C., & Lyu, R. Y. (2004). A Taiwanese Text-to-Speech System with Application to Language Learning, in *Proceedings of the IEEE International Conference on Advanced Learning Technologies, 2004.*

Lin, C. J. & Chen, H. H. (1999). A Mandarin to Taiwanese Min Nan Machine Translation System with Speech Synthesis of Taiwanese Min Nan, *International Journal of Computational Linguistics and Chinese Language Processing*, 14(1), 59-84.

Lin, Y. C. (2006). *The Prediction of Pronunciation of Polyphonic Characters in a Mandarin Text-to-Speech System*, Master thesis, Department of Computer Science and Engineering, National Chung Hsing University.

Lin, Y. J. & Yu, M. S. (1998). An Efficient Mandarin Text-to-Speech System on Time Domain, *IEICE Transactions on Information and Systems*, E81-D(6), June 1998, 545-555.

Lin, Y. J., Yu, M. S., Lin, C. Y., & Lin, Y. T. (2008). A Multi-Layered Approach to the Polysemy Problems in a Chinese to Taiwanese TTS System, in *Proceeding of 2008 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, June, 2008, 428-435.

Lu, H. M. (2002). *An Implementation and Analysis of Mandarin Speech Synthesis Technologies*, M. S. Thesis, Institute of Communication Engineering, National Chiao-Tung University, June 2002.

Pan, N. H. & Yu, M. S. (2008). Improving Intonation Modules in Chinese TTS Systems, in *The 13th Conference on Artificial Intelligence and Applications (TAAI 2008)*, 329-336, Nov. 21-22, 2008, Yilan, Taiwan.

Pan, N. H., Yu, M. S., & Tsai, C. M. (2008). A Mandarin Text to Taiwanese Speech System, in *The 13th Conference on Artificial Intelligence and Applications (TAAI 2008)*, 1-5, Nov. 21-22, 2008, Yilan, Taiwan.

Shih, C. & Sproat, R. (1996). Issues in Text-to-Speech Conversion for Mandarin, *Computational Linguistics and Chinese Language Processing*, 1(1), 37-86.

Wu, C. H., Hsia, C. C., Chen, J. F., & Wang, J. F. (2007). Variable-Length Unit Selection in TTS Using Structural Syntactic Cost, *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4), 1227-1235.

Yang, Y. C. (1999). *An Implementation of Taiwanese Text-to-Speech System*, Master thesis, Department of Communication Engineering, National Chiao Tung University, 1999.

Yu, M. S., Chang, T. Y., Hsu, C. H., & Tsai, Y. H. (2005). A Mandarin Text-to-Speech System Using Prosodic Hierarchy and a Large Number of Words, in *Proc. 17th Conference on Computational Linguistics and Speech Processing, (ROCLING XVII)*, 183-202, Sep. 15-16, 2005, Tainan, Taiwan.

Yu, M. S. & Huang, F. L. (2003). Disambiguating the Senses of Non-Text Symbols for Mandarin TTS Systems with a Three-Layer Classifier, *Speech Communication*, 39(3-4), 191-229.

Zhong, X. R. (1999). *An Improvement on the Implementation of Taiwanese TTS System*, Master thesis, Department of Communication Engineering, National Chiao Tung University.